

# Mine Classification With Imbalanced Data

David P. Williams, *Member, IEEE*, Vincent Myers, and Miranda Schatten Silvius

**Abstract**—In many remote-sensing classification problems, the number of targets (e.g., mines) present is very small compared with the number of clutter objects. Traditional classification approaches usually ignore this class imbalance, causing performance to suffer accordingly. In contrast, the recently developed infinitely imbalanced logistic regression (IILR) algorithm explicitly addresses class imbalance in its formulation. We describe this algorithm and give the details necessary to employ it for remote-sensing data sets that are characterized by class imbalance. The method is applied to the problem of mine classification on three real measured data sets. Specifically, classification performance using the IILR algorithm is shown to exceed that of a standard logistic regression approach on two land-mine data sets collected with a ground-penetrating radar and on one underwater-mine data set collected with a sidescan sonar.

**Index Terms**—Classification, imbalanced data, land mines, logistic regression (LR), mine detection, radar, sonar, underwater mines.

## I. INTRODUCTION

IN MANY remote-sensing classification problems, the number of targets present is very small compared with the number of clutter objects encountered. For example, in land-mine detection applications, it is common to have nearly 100 false alarms due to clutter for every real mine present. Similarly, in underwater-mine classification applications, the number of naturally occurring clutter objects (such as rocks) that are detected typically far outweighs the relatively rare event of detecting a mine.

Nevertheless, such class imbalance is often implicitly ignored when it comes time to choose a classification algorithm to employ. As a result, the traditional classification approaches that do not account for severe class imbalance often lead to poor classification performance.

A recently developed classification technique that is named infinitely imbalanced logistic regression (IILR) [1] explicitly acknowledges the problem of class imbalance in its formulation. Although the method was developed with an eye toward applications involving rare events such as fraud detection or drug discovery, we apply the technique to the problem of mine classification in remote-sensing data. Specifically, we demonstrate the benefits of the proposed approach on three data sets of real measured remote-sensing data.

Manuscript received January 28, 2009; revised March 4, 2009 and April 7, 2009. First published June 10, 2009; current version published July 4, 2009.

D. P. Williams is with the NATO Undersea Research Centre, 19126 La Spezia, Italy (e-mail: williams@nurc.nato.int).

V. Myers is with the Defence R&D Canada, Halifax, NS B2Y 3Z7, Canada (e-mail: vincent.myers@drdc-rddc.gc.ca).

M. S. Silvius is with the U.S. Army Night Vision & Electronic Sensors Directorate, Fort Belvoir, VA 22060 USA (e-mail: miranda.silvius@us.army.mil).

Digital Object Identifier 10.1109/LGRS.2009.2021964

Various approaches have been developed to handle the issue of class imbalance in general. However, despite their relevance for many applications in the remote-sensing community, such methods have not been widely adopted. One notable exception to this is [2], which sought to classify agricultural crops in multispectral imagery by employing neural networks that were specially designed for class imbalance. A second example is [3], which used an *ad hoc* method for detecting oil spills in synthetic aperture sonar imagery. This letter is the first to address the issue of class imbalance for mine classification.

The remainder of this letter is organized in the following manner. In Section II, we describe the IILR algorithm and also provide the necessary implementation details for its successful application to classification problems. In Section III, we describe the three real measured remote-sensing data sets—two land-mine data sets from a ground-penetrating radar (GPR), and one underwater-mine data set from a sidescan sonar—on which we apply the IILR algorithm. Experimental classification results for the three data sets are shown in Section IV, before concluding remarks are made in Section V.

## II. CLASSIFICATION

### A. LR

Logistic regression (LR) is a commonly used approach for performing binary classification. It learns a set of parameters,  $\{w_0, \mathbf{w}\}$ , that maximizes the likelihood of the class labels for a given set of training data.

Let  $\mathbf{x}_i \in \mathbb{R}^d$  denote a (column) vector of  $d$  features representing the  $i$ th data point, and  $y_i \in \{0, 1\}$  denote its corresponding class label (e.g., clutter or mine). For a labeled (training) data point,  $y_i$  is known; for an unlabeled (testing) data point,  $y_i$  is unknown.

Under the LR model, the probability of label  $y_i = 1$  given  $\mathbf{x}_i$  is

$$\psi_i \equiv p(y_i = 1 | \mathbf{x}_i) = \frac{\exp\{w_0 + \mathbf{w}^T \mathbf{x}_i\}}{1 + \exp\{w_0 + \mathbf{w}^T \mathbf{x}_i\}} \quad (1)$$

where  $w_0 \in \mathbb{R}$  and  $\mathbf{w} \in \mathbb{R}^d$  are the LR intercept and coefficients, respectively. For a set of  $N$  independent labeled data points,  $\{\mathbf{x}_i, y_i\}_{i=1}^N$ , the log-likelihood of the class labels can be written as

$$\ell(w_0, \mathbf{w}) = \sum_{i=1}^N [(1 - y_i) \log(1 - \psi_i) + y_i \log \psi_i]. \quad (2)$$

To maximize the log-likelihood in (2), a standard optimization approach can be employed, since the gradient (and Hessian) of (2) with respect to  $\{w_0, \mathbf{w}\}$  can be readily

calculated. Once the LR parameters  $\{w_0, \mathbf{w}\}$  have been learned, the probability that an unlabeled testing data point  $\mathbf{x}_i$  belongs to each class can be obtained via (1).

### B. IILR

When the number of data points belonging to one class far exceeds the number belonging to the other class, the standard LR approach can lead to poor classification performance. This scenario of class imbalance motivates the use of a modified form of LR that was introduced in [1] named IILR.

Let  $N_1$  and  $N_0$  be the number of labeled data points belonging to classes  $y = 1$  and  $y = 0$ , respectively. Let the  $i$ th labeled data point belonging to class  $y = 1$  be written as  $\mathbf{x}_{1i}$ ; similarly, write the  $i$ th labeled data point belonging to class  $y = 0$  as  $\mathbf{x}_{0i}$ . With this notation, the log-likelihood (2) can be rewritten as

$$\begin{aligned} \ell(w_0, \mathbf{w}) = & - \sum_{i=1}^{N_0} [\log(1 + \exp\{w_0 + \mathbf{w}^T \mathbf{x}_{0i}\})] \\ & + \sum_{i=1}^{N_1} [w_0 + \mathbf{w}^T \mathbf{x}_{1i} - \log(1 + \exp\{w_0 + \mathbf{w}^T \mathbf{x}_{1i}\})]. \end{aligned} \quad (3)$$

Without loss of generality, assume that class  $y = 1$  is the rare class so that  $N_1 \ll N_0$ . Also, assume that  $N_0$  is very large so that replacing the first sum in (3) by an integral is justifiable. Then, write the average of the class  $y = 1$  data points as

$$\bar{\mathbf{x}} = \frac{1}{N_1} \sum_{i=1}^{N_1} \mathbf{x}_{1i}. \quad (4)$$

With these changes and additional algebraic manipulation, (3) can be rewritten as

$$\begin{aligned} \ell(w_0, \mathbf{w}) = & N_1 w_0 - \sum_{i=1}^{N_1} \log(1 + \exp\{w_0 + \mathbf{w}^T (\mathbf{x}_{1i} - \bar{\mathbf{x}})\}) \\ & - N_0 \int \log(1 + \exp\{w_0 + \mathbf{w}^T (\mathbf{x} - \bar{\mathbf{x}})\}) p(\mathbf{x}|y=0) d\mathbf{x}. \end{aligned} \quad (5)$$

The major insight discovered in [1] is that as  $N_0 \rightarrow \infty$  (while  $N_1$  remains fixed), the LR intercept  $w_0 \rightarrow -\infty$ , but the coefficients  $\mathbf{w}$  do not necessarily diverge. Specifically, the limiting value of  $\mathbf{w}$  that maximizes (5) is found to satisfy

$$\bar{\mathbf{x}} = \frac{\int \mathbf{x} \exp\{\mathbf{w}^T \mathbf{x}\} p(\mathbf{x}|y=0) d\mathbf{x}}{\int \exp\{\mathbf{w}^T \mathbf{x}\} p(\mathbf{x}|y=0) d\mathbf{x}}. \quad (6)$$

It is also demonstrated in [1] that when the conditional distribution of the data  $\mathbf{x}$  given that  $y = 0$ ,  $p(\mathbf{x}|y=0)$  is either a Gaussian or a mixture of Gaussians,  $\mathbf{w}$  is readily obtainable.

1) *Classifier Learning*: For the case of a single Gaussian (with mean  $\boldsymbol{\mu}$  and covariance  $\boldsymbol{\Sigma}$ )

$$p(\mathbf{x}|y=0) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad (7)$$

the limiting solution for the (infinitely imbalanced) LR coefficients can be written analytically as [1]

$$\mathbf{w} = \boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}). \quad (8)$$

For the case of a  $K$ -component Gaussian mixture model (GMM)

$$p(\mathbf{x}|y=0) = \sum_{k=1}^K \alpha_k \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \quad (9)$$

(where  $\alpha_k$ ,  $\boldsymbol{\mu}_k$ , and  $\boldsymbol{\Sigma}_k$  are the mixing proportion, mean, and covariance, respectively, of the  $k$ th Gaussian), the limiting solution for the (infinitely imbalanced) LR coefficients is the solution to [1]

$$\sum_{k=1}^K \alpha_k (\boldsymbol{\mu}_k + \boldsymbol{\Sigma}_k \mathbf{w} - \bar{\mathbf{x}}) \exp\{\mathbf{w}^T \boldsymbol{\mu}_k + \mathbf{w}^T \boldsymbol{\Sigma}_k \mathbf{w} / 2\} = 0. \quad (10)$$

Solving for  $\mathbf{w}$  when  $p(\mathbf{x}|y=0)$  is modeled as a  $K$ -component GMM can be cast as a convex optimization problem. Specifically, the objective function to be minimized, its gradient, and its Hessian are given as [1]

$$\mathbf{f}(\mathbf{w}) = \sum_{k=1}^K \alpha_k \exp\{\mathbf{w}^T (\boldsymbol{\mu}_k - \bar{\mathbf{x}}) + \mathbf{w}^T \boldsymbol{\Sigma}_k \mathbf{w} / 2\} \quad (11)$$

$$\begin{aligned} \mathbf{g}(\mathbf{w}) = & \sum_{k=1}^K \alpha_k (\boldsymbol{\mu}_k + \boldsymbol{\Sigma}_k \mathbf{w} - \bar{\mathbf{x}}) \\ & \times \exp\{\mathbf{w}^T (\boldsymbol{\mu}_k - \bar{\mathbf{x}}) + \mathbf{w}^T \boldsymbol{\Sigma}_k \mathbf{w} / 2\} \end{aligned} \quad (12)$$

$$\begin{aligned} \mathbf{H}(\mathbf{w}) = & \sum_{k=1}^K \alpha_k (\boldsymbol{\Sigma}_k + (\boldsymbol{\mu}_k + \boldsymbol{\Sigma}_k \mathbf{w} - \bar{\mathbf{x}})(\boldsymbol{\mu}_k + \boldsymbol{\Sigma}_k \mathbf{w} - \bar{\mathbf{x}})^T) \\ & \times \exp\{\mathbf{w}^T (\boldsymbol{\mu}_k - \bar{\mathbf{x}}) + \mathbf{w}^T \boldsymbol{\Sigma}_k \mathbf{w} / 2\} \end{aligned} \quad (13)$$

respectively.

In this work, Newton's method was employed to solve for  $\mathbf{w}$ . For the sake of completeness, we provide the iterative formula of the method here, which is given by

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \left( \mathbf{H}(\mathbf{w}^{(t)}) \right)^{-1} \mathbf{g}(\mathbf{w}^{(t)}) \quad (14)$$

where  $\mathbf{w}^{(t)}$  is the set of parameters at iteration  $t$ .

2) *Classifier Prediction*: In the infinitely imbalanced limit where  $N_0 \rightarrow \infty$ ,  $\mathbf{w}$  converges to a finite limit while  $w_0 \rightarrow -\infty$  [1]. Therefore, the posterior probability that an unlabeled data point belongs to class  $y = 1$  using (1) would result in  $p(y = 1|\mathbf{x}_i) \rightarrow 0$  for all  $\mathbf{x}_i$ .

Instead of using (1) to rank the unlabeled data points (e.g., from most mine-like to least mine-like), one can instead use the quantity

$$\psi_i = \mathbf{w}^T \mathbf{x}_i \quad (15)$$

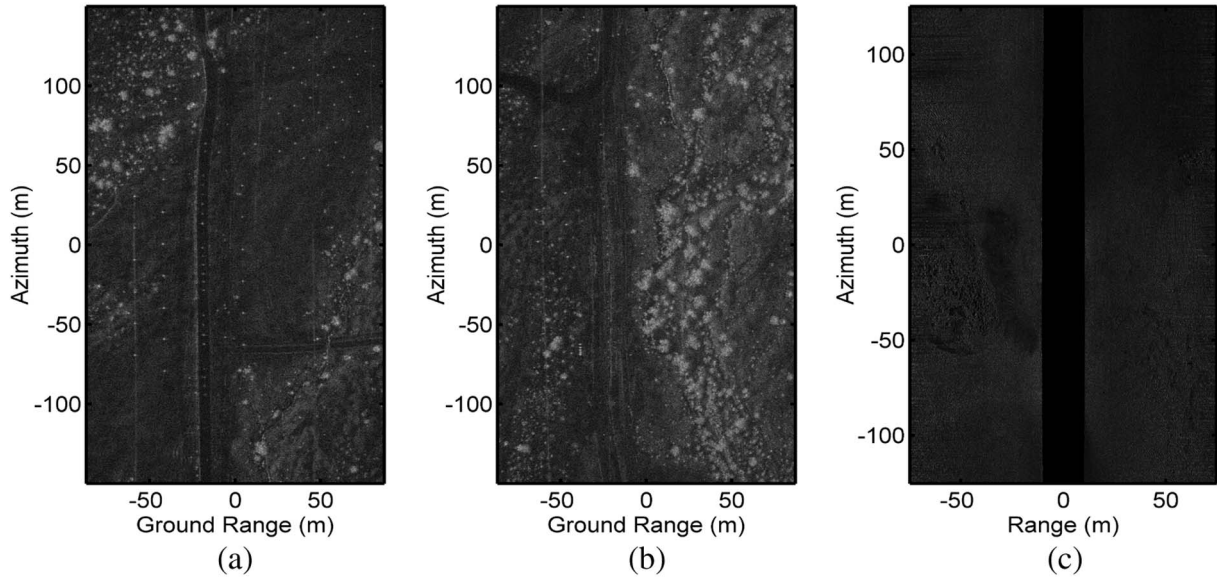


Fig. 1. Raw-magnitude images from the (a) land mine A and (b) land mine B data sets. (c) An image swath from the underwater-mine data set.

to do so. Thus, after the IILR parameters  $w$  have been learned, any unlabeled testing data point can be evaluated using (15).

3) *GMM Estimation*: Central to the IILR algorithm is the ability to model the conditional distribution of the data  $x$  given that  $y = 0$ ,  $p(x|y = 0)$ , as a mixture of Gaussians. Because of its prime importance, we devote particular attention to this estimation.

It is well known that a GMM can accurately model an arbitrary continuous distribution. Moreover, this model estimation can efficiently be performed using the expectation-maximization (EM) algorithm [4] or the variational Bayesian EM (VB-EM) algorithm [5]. Whereas the result of the EM algorithm will be point estimates for the parameters (i.e., the mixing proportions, means, and covariances), the result of the VB-EM algorithm will be full *distributions* of the parameters.

We choose to employ the VB-EM algorithm for GMM estimation in this letter because with an EM approach, one must specify the number of mixture components *a priori*. In contrast, the VB-EM approach provides a principled way, via the *evidence* (i.e., marginal likelihood) [6], to establish the appropriate number of mixture components that are represented by the data. Whereas the standard EM algorithm will always favor increasingly complex models (i.e., more mixture components), the variational formulation automatically penalizes overly complex models.

The equations for learning the posterior distribution of the GMM parameters via the VB-EM algorithm are given in [5], so we do not repeat them here.

### III. DATA SETS

#### A. GPR Data for Land-Mine Classification

We consider two land-mine data sets containing buried targets at a U.S. Army desert test site. The data were collected by a Mirage GPR sensor located on an airborne platform. The sensor is a 300-MHz to 3-GHz system with polarimetric capability. The resulting image resolution was 11.7 cm in ground range

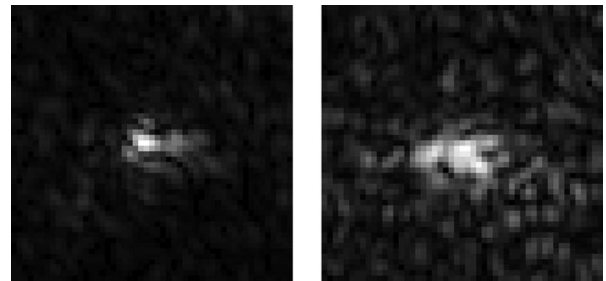


Fig. 2. Two typical chips from the land-mine data sets.

and 4.77 cm in azimuth. The two GPR (magnitude) images are shown in Fig. 1(a) and (b).

Each of the two images undergoes the same preprocessing steps. However, it should be emphasized that because the preprocessing procedure is not the main focus of this letter, we do not describe the procedure in great detail.

1) *Detection*: The detection stage proceeded as follows. First, the raw-magnitude image was smoothed and dilated. Next, the gradient of this blurred image was computed in both the vertical and horizontal directions to form a gradient image. A filter template that mimics the response of a mine was then correlated with the gradient image to produce a filtered image. The highest “energy” locations in this filtered image were recorded as alarms to be investigated further.

It has been observed that this detection procedure effectively rejects many false alarms due to vegetation that a simple energy detector applied to the raw-magnitude image would flag as alarms.

2) *Feature Extraction*: For each alarm from the detection stage, a smaller image chip containing the object was then extracted from the original large raw image. Two typical representative chips from the data set are shown in Fig. 2.

Seven features were subsequently extracted for each alarm. These features are the “energy” score from the detection stage, four moment-based features of the pixel values of the magnitude chip (namely, the mean, variance, skewness, and kurtosis),

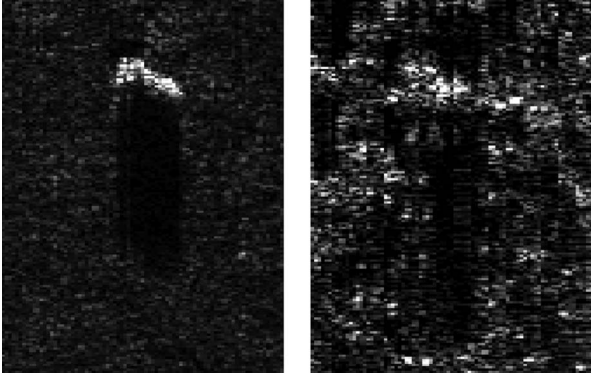


Fig. 3. Two typical chips from the underwater-mine data set.

a measure of the spatial variance in the magnitude chip, and the entropy of the pixel values in the corresponding phase chip.

At the end of this preprocessing, each alarm is represented by a vector of seven features. Classification is performed on these data.

#### B. Sidescan Sonar Data for Underwater-Mine Classification

We also consider one underwater-mine data set from a site in the Gulf of Mexico, near Panama City, Florida, U.S. The data were collected as part of The Technical Cooperation Program (TTCP) trial *Mongoose* by a Klein 5500 sidescan sonar, which was towed by the Canadian Interim Remote Minehunting and Disposal System. The sensor has a center frequency of 455 kHz and a bandwidth of 20 kHz. The resulting image resolution was 3.29 cm in range and approximately 10 cm in azimuth. One image swath from the data set is shown in Fig. 1(c).

Ground truth information is available because targets were deployed at the site. The naturally occurring clutter includes large rocks and seabed characteristics that produce target-like signatures.

1) *Detection*: The detection procedure followed the approach described in [7] in which four simple detectors are applied to the sonar imagery. Two of the detectors employ matched filters that search for an “ideal” target signature, whereas the other two calculate certain statistical quantities indicative of targets. The alarms generated by the four detectors were pooled and investigated further.

2) *Feature Extraction*: For each alarm from the detection stage, a smaller image chip containing the object was then extracted from the original swath image. Two typical representative chips from the data set are shown in Fig. 3.

Subsequently, each chip was segmented into three regions—object echo, object shadow, and background—using the approach described in [8]. Twenty-two features related to the geometrical or statistical quantities of the echo and shadow were then extracted from each segmented chip. The features are a subset of those used in [9]; some example features include the angle and the ratio of the major and minor axes of an ellipse fit to the segmented pixels, the statistics of the height profile of the object, the length and width of the object, and a measure of the convexity of the segmented areas.

At the end of this preprocessing, each alarm is represented by a vector of 22 features. Classification is performed on these data.

TABLE I  
DETAILS OF THE THREE DATA SETS

DATA SET	NUMBER OF FEATURES	NUMBER OF DATA POINTS	
		TARGETS	CLUTTER
LAND MINE A	7	22	798
LAND MINE B	7	11	867
UNDERWATER MINE	22	16	2650

## IV. EXPERIMENTAL RESULTS

### A. Experimental Set-Up

The details of the three data sets relevant for the classification stage are summarized in Table I.

Three different classification methods are applied to each of the three data sets. The baseline approach is a standard LR algorithm that does not account for class imbalance. The proposed IILR algorithm is also applied. Two different versions of this method are considered, which differ only in the number of Gaussians that are used to model the distribution of the clutter class of data. The first version models the distribution of the clutter class of data using a single Gaussian ( $K = 1$ ), whereas the second version uses a GMM with a maximum of  $K = 10$  Gaussians.

For each of the three data sets, the same stratified five-fold cross-validation training and testing procedure is employed. Specifically, the cross-validation was stratified in the sense that the proportion of samples from each class is approximately equal in each of the five folds. (With the imbalanced data sets considered, this means that each fold contains only two to five targets.) The data points in one of the folds are treated as unlabeled testing data, whereas the data points in the other folds are treated as labeled training data. The classifier is learned using the training data, and is then exploited to classify the data points in the testing fold. This process is then repeated so that each fold is treated as unlabeled testing data one time.

For each classification method, the output for a given unlabeled data point is related to the probability of belonging to the mine class. All data points for which this quantity is greater than or equal to some threshold  $\tau$  are classified as targets (i.e., mines), whereas the other data points are classified as clutter. Varying the threshold  $\tau$  will then effectively generate a receiver operating characteristic (ROC) curve.

The area under the ROC curve (AUC), which provides a scalar summary measure of performance, can also be easily calculated. The AUC is given by the Wilcoxon statistic [10]

$$\text{AUC} = \frac{1}{N_1 N_0} \sum_{i=1}^{N_1} \sum_{j=1}^{N_0} \mathbb{I}_{\psi_{1i} > \psi_{0j}} \quad (16)$$

where  $\psi_{1i}$  are the classification scores of the data belonging to the target class,  $\psi_{0j}$  are the classification scores of the data belonging to the clutter class, and  $\mathbb{I}$  is an indicator function. (The “classification score” of a given data point used in the AUC calculation for the LR and IILR methods is the quantity given in (1) and (15), respectively.)

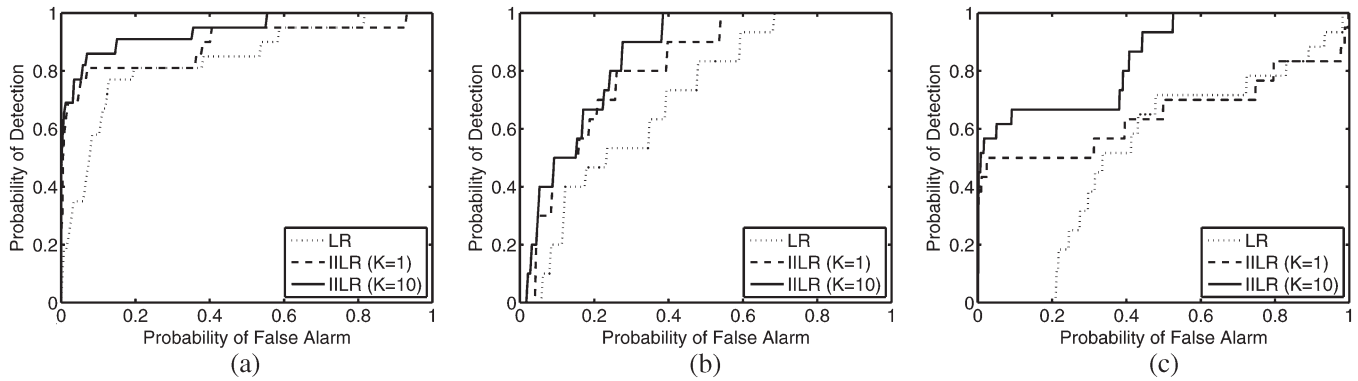


Fig. 4. Average ROC curves of the classification methods when using stratified five-fold cross-validation for the (a) land mine A, (b) land mine B, and (c) underwater-mine data sets.

TABLE II  
AUC (MEAN  $\pm$  1 STANDARD DEVIATION) OF THE CLASSIFICATION METHODS WHEN USING STRATIFIED FIVE-FOLD CROSS-VALIDATION

DATA SET	LR	IILR	
		( $K = 1$ )	( $K = 10$ )
LAND MINE A	$0.8367 \pm 0.0787$	$0.8906 \pm 0.0791$	$0.9417 \pm 0.0513$
LAND MINE B	$0.7087 \pm 0.1287$	$0.8137 \pm 0.1716$	$0.8529 \pm 0.1245$
UNDERWATER MINE	$0.5286 \pm 0.1634$	$0.6497 \pm 0.2019$	$0.8480 \pm 0.1393$

### B. Mine-Classification Results

The mean classification performance for the three different classification methods on each of the three data sets considered is shown in terms of ROC curves in Fig. 4. The classification performance in terms of AUC scores for the three data sets is compactly summarized in Table II.

As can be seen from Fig. 4 and Table II, for all the three data sets considered, the IILR approach that accounts for class imbalance outperforms the standard LR approach. (Additional results in terms of AUC using a leave-one-out procedure rather than the stratified five-fold cross-validation, which are not shown here due to space constraints, also support the claim that the IILR approach is superior to the LR approach.)

These results were made possible by the severe class imbalance of the data sets, in which the numbers of clutter data points far outweighed the numbers of target data points. However, a less obvious reason for the success of the IILR approach is the fact that the features of the target data points were well represented by their mean values,  $\bar{x}$ . In general, the IILR approach will not be suitable for data sets in which the data points of the rare class do not cluster tightly.

## V. CONCLUSION

In this letter, a recently developed classification approach for data sets characterized by class imbalance has been described. The implementation details needed to employ the approach have also been provided. The utility of the IILR algorithm for

mine classification was demonstrated on three real measured remote-sensing data sets.

Many classification algorithms in use today are relatively similar to LR. Therefore, after introducing this method to the community, we expect that other researchers will find the approach useful for other applications that are characterized by severe class imbalance (such as unexploded ordnance detection [11], [12]).

## REFERENCES

- [1] A. Owen, "Infinitely imbalanced logistic regression," *J. Mach. Learn. Res.*, vol. 8, pp. 761–773, May 2007.
- [2] L. Bruzzone and S. Serpico, "Classification of imbalanced remote-sensing data by neural networks," *Pattern Recognit. Lett.*, vol. 18, no. 11–13, pp. 1323–1328, Nov. 1997.
- [3] M. Kubat, R. Holte, and S. Matwin, "Machine learning for the detection of oil spills in satellite radar images," *Mach. Learn.*, vol. 30, no. 2/3, pp. 195–215, Feb./Mar. 1998.
- [4] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Stat. Soc., B*, vol. 39, no. 1, pp. 1–38, 1977.
- [5] N. Nasios and A. Bors, "Variational expectation-maximization training for Gaussian networks," in *Proc. IEEE Workshop Neural Netw. Signal Process.*, 2003, pp. 339–348.
- [6] M. Beal and Z. Ghahramani, "The variational Bayesian EM algorithm for incomplete data: Application to scoring graphical model structures," *Bayesian Stat.*, vol. 7, pp. 453–464, 2003.
- [7] V. Myers and Ø. Midtgaard, "Fusion of contacts in synthetic aperture sonar imagery using performance estimates," in *IOA Int. Conf. Detect. Classif. Underwater Targets*, 2007, pp. 77–88.
- [8] V. Myers, "Sonar image segmentation using iteration and fuzzy logic," in *Proc. CAD/CAC Conf.*, 2001.
- [9] J. Fawcett and V. Myers, "Computer-aided classification for a database of images of minelike objects," Defence R&D Canada-Atlantic, Dartmouth, NS, Canada, Tech. Rep. DRDC Atlantic TM 2004-272, 2005.
- [10] J. Hanley and B. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, pp. 29–36, 1982.
- [11] L. Beran and D. Oldenburg, "Selecting a discrimination algorithm for unexploded ordnance remediation," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 9, pp. 2547–2557, Sep. 2008.
- [12] Q. Liu, X. Liao, and L. Carin, "Detection of unexploded ordnance via efficient semi-supervised and active learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 9, pp. 2558–2567, Sep. 2008.